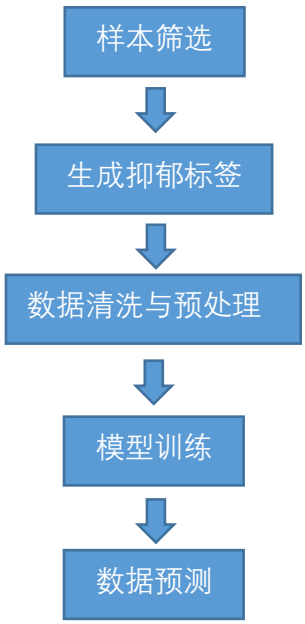
1. 流程描述

本项目旨在对 60 岁及以上的老年人是否抑郁做出预测。目前现有的全年龄样本量有 20948 个，过滤之后 60 岁及以上的样本量有一万多个。我们基于这一万多个的样本量构建数据集，同时根据 CES-D 表的得分来区分样本抑郁与否（总分 30 分，大于等于 12 分为抑郁，标签为 1，小于 12 为非抑郁，标签为 0）。接着训练模型，用训练好的模型对 2015 及 2018 年的样本数据进行预测，并通过比较预测结果与真实结果来衡量模型效果。



2. 数据清洗与预处理

对从原始数据库抽取的数据进行清洗与预处理，具体包括：

①重复样本剔除

②缺失值填充

③独立热编码（用于将类别变量转化为数值变量）

④数据标准化 （对于消除不同变量之间的量纲差异以及增强分类模型的鲁棒性）

其中在缺失值填充阶段，我们考虑采用 KNN 算法进行缺失值填充，它通过计算高维空间上样本之间的距离来识别相邻点，并利用相邻观测值的完整值来估计缺失值，相比以往常用的均值及中位数填充方法更为可靠。

3. 类别不平衡问题的处理

对于在实际建模过程中存在的目标变量类别不平衡问题(即样本中非抑郁人群数量远超抑郁人群数量，导致二个类别比例极不均衡)。我们拟采用：

① 上采样方法（从小样本类别中随机抽取一定量的样本来扩充样本量，使得与大样本量类别比例均衡）

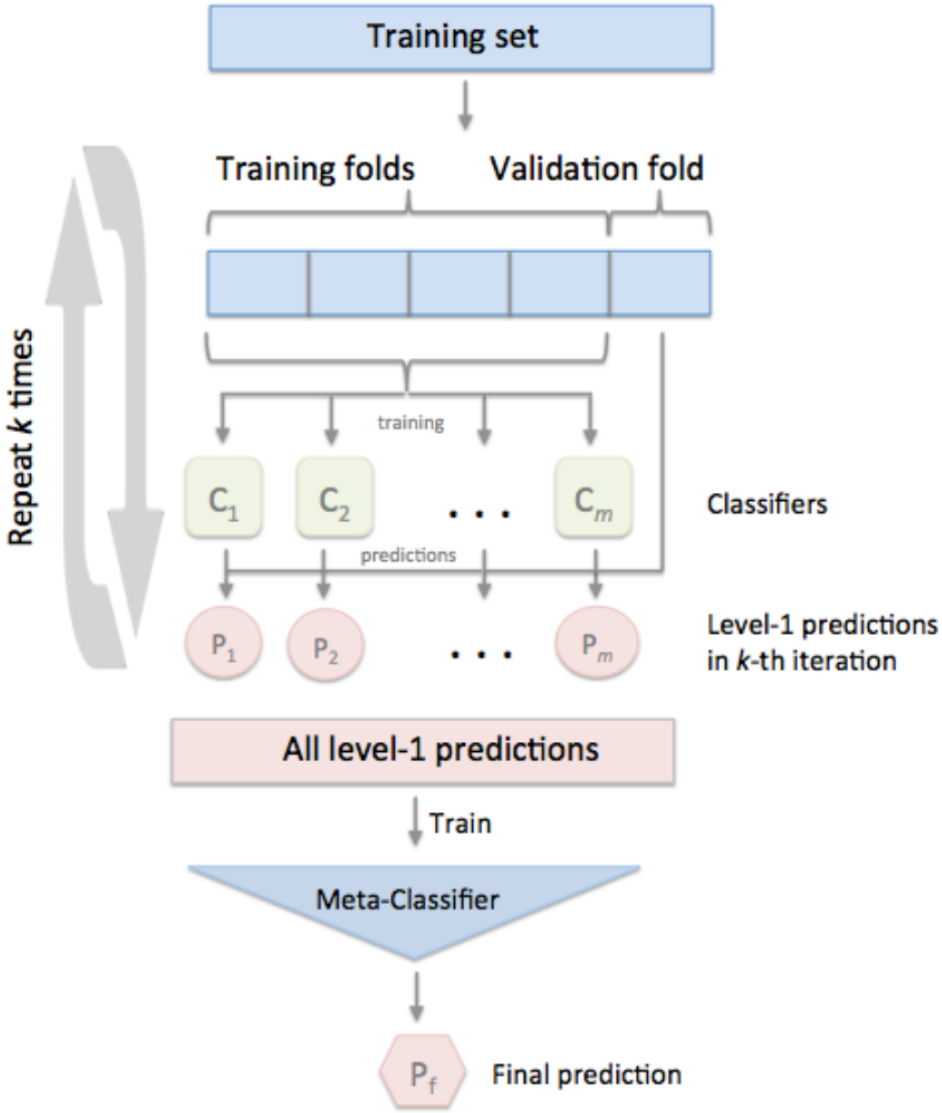
②下采样方法（从大样本类别中随机抽取一定量的样本，使得抽取的样本量与小样本量类别比例均衡）

③SMOTE（Synthetic Minority Over-sampling Technique）方法，即合成少数类过采样技术，它是基于随机过采样算法的一种改进方案，其基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中来达到均衡样本类别的目的。

4. 模型构建

在模型构建阶段，我们拟采用 Model Stacking 的集成方法来训练数据。Model Stacking 方法的基本思想是同时训练多个算法原理不尽相同的机器学习算法，其中每个模型都包含特征工程、特征选择、超参数调优等必要步骤，并在此基础上训练一个元模型来组合它们，然后基于这些弱模型返回的多个预测结果输出最终的预测结果。同时，我们在融合算法的构建过程中采用 K 折交叉训练的方式来降低发生过拟合的可能性。

Model Stacking 方法的说明图请参考下图：



5. 模型评估

   我们拟采用 Accuracy、AUC、Precision 等指标来分别对单个模型以及融合模型进行评估。

6. 特征重要性

   在我们选择 tree-based 类机器学习模型训练数据集时（诸如随机森林，Xgboost 等）,我们可以通过计算基尼重要性来排列出具有较高重要性的特征变量。

   另外，我们拟采用 SHAP 值来计算特征重要性，SHAP 值是通过博弈论中的 Shapley 值来估计每个特征如何对预测作出贡献。SHAP 值相比于基尼重要性能提供更多的信息，如特征如何会如何影响预测值等。

   我们希望通过计算特征重要性来指导我们设计问卷过程中需要注意的重要问题。